URBAN SOUND CLASSIFICATION

MARC KELECHAVA

PROBLEM OVERVIEW

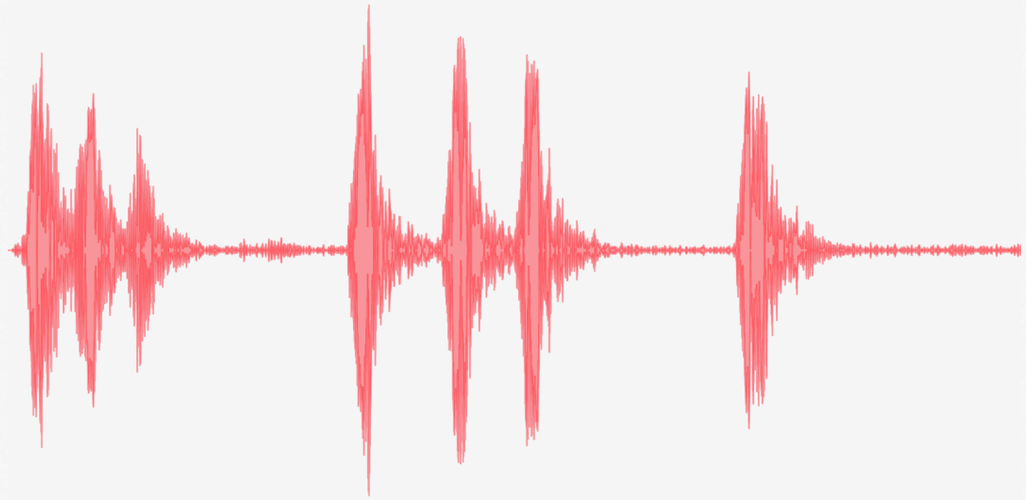
URBAN AUDIO CLASSIFICATION

Example business case: auto-detecting excess noise in high-complaint areas for a city government.

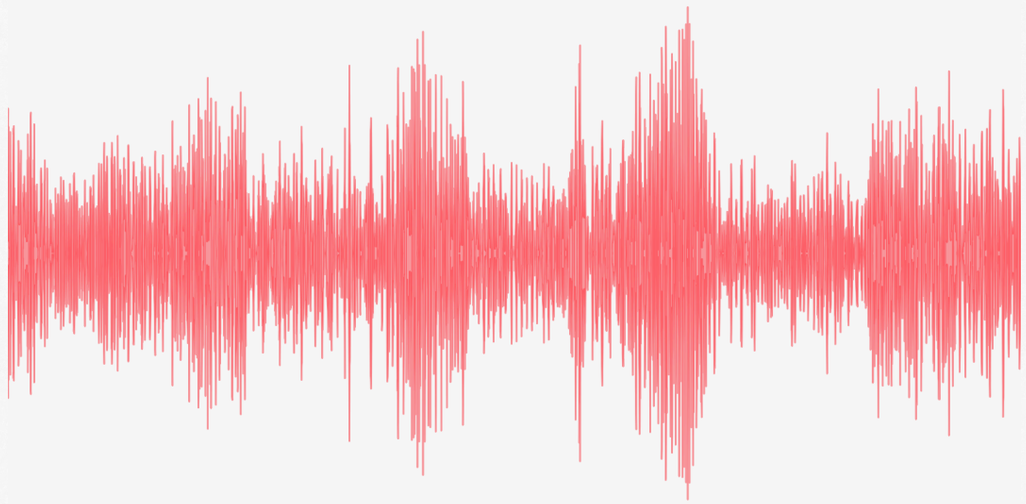
To start: I took 8,732 4-second audio clips of live urban recordings in 10 classes: e.g. Jackhammers, Dog Barks, etc.

18 hours of raw audio in total.

DOG AUDIO WAVEFORM



STREET MUSIC WAVEFORM



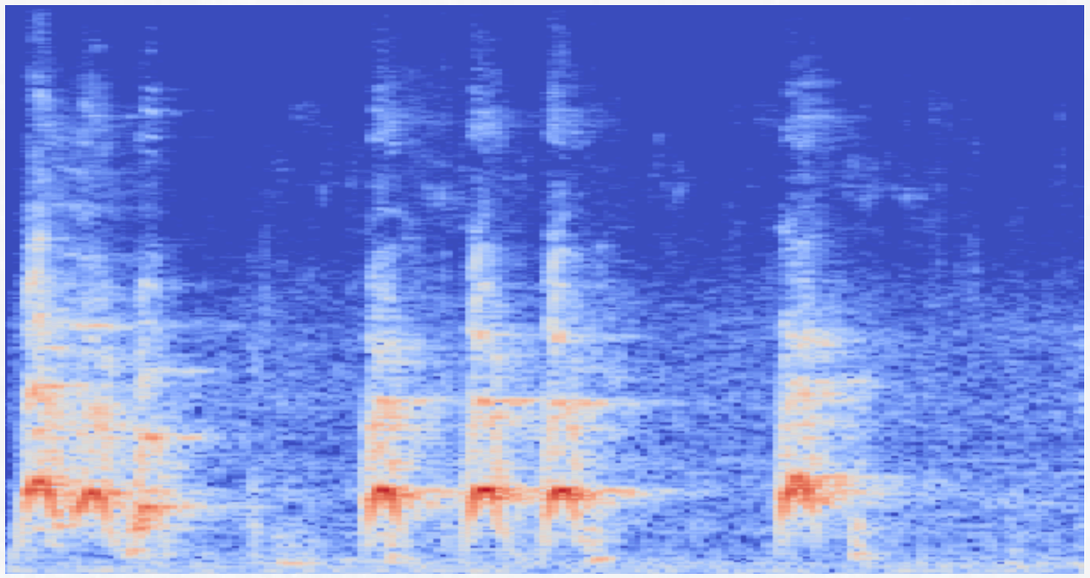
DATA SOURCE:

J. Salamon, C. Jacoby and J. P. Bello, "A Dataset and Taxonomy for Urban Sound Research".

AUDIO FEATURE EXTRACTION

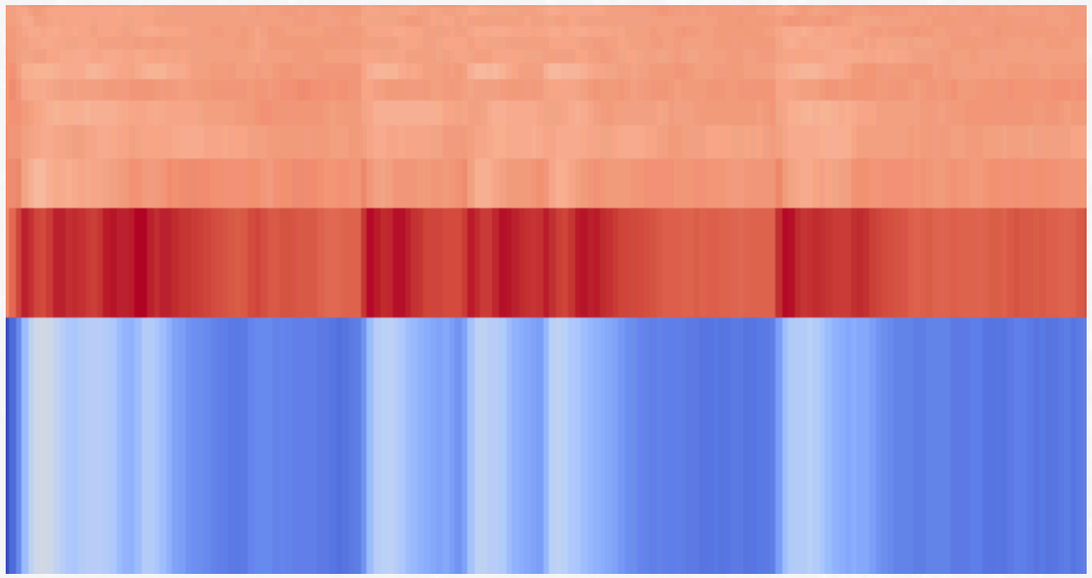
FROM THOUSANDS OF FEATURES TO 66-DIMENSIONS

DOG AUDIO 'VOICEPRINT'



3 BARKS 3 BARKS 1 BARK

REDUCED FEATURE SET



3 BARKS 3 BARKS 1 BARK

Began with (8000+) 88,200-vectors using 22,050 samples a second for 4 seconds in each audio clip.

AWS XXL SERVER FOR CONVERSION SPEED

10 ‘GROUPS’ OF CLIPS FROM 10 LONG RECORDINGS

DANGER OF LEAKING TEST DATA DUE TO RECORDING ‘METADATA’

LEAVE ONE GROUP OUT

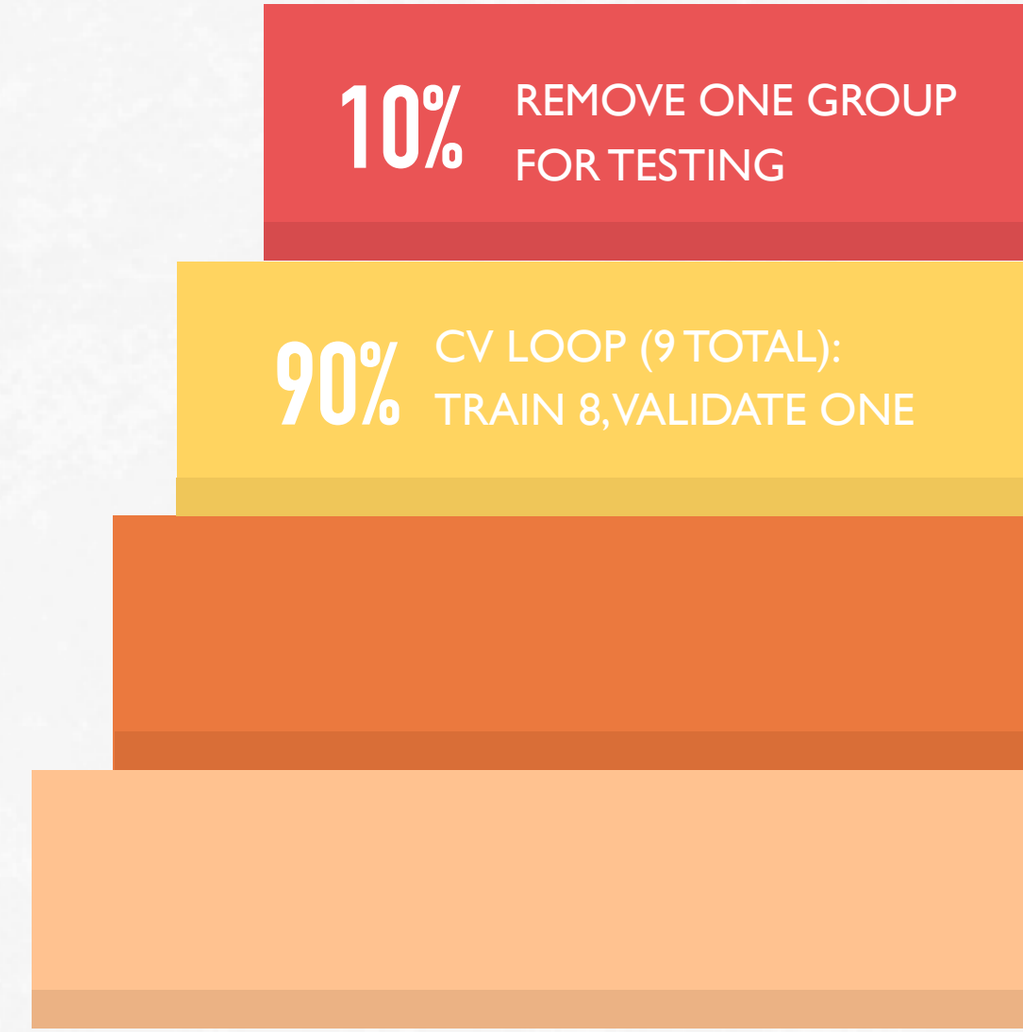
CROSS VALIDATION TECHNIQUE

1. The 'Leave One Group Out' SKLearn function lets you pass in group 'membership' information.

   With very careful preprocessing you can make sure the model is not trained on elements of the 'base' recording.

2. 10 base 'groups' from 10 long recordings. Take one for test, 9 for CV.

   Run LeaveOneGroupOut CV on the 9.



LEAVE ONE GROUP OUT

CROSS VALIDATION TECHNIQUE

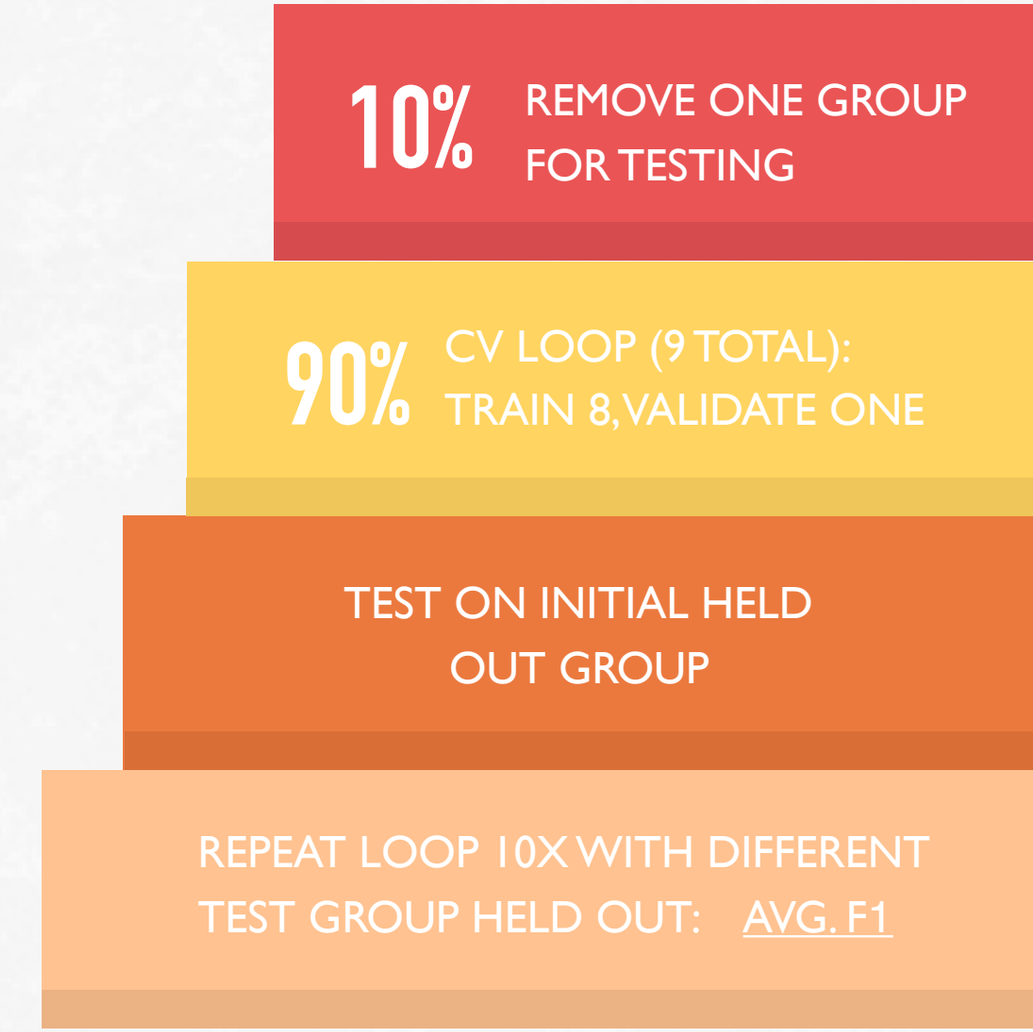
1. The 'Leave One Group Out' SKLearn function lets you pass in group 'membership' information.

   With very careful preprocessing you can make sure the model is not trained on elements of the 'base' recording.

2. 10 base 'groups' from 10 long recordings. Take one for test, 9 for CV.

   Run LeaveOneGroupOut CV on the 9.

   Within same outer loop: score on test holdout. Repeat 10 times with different test groups and average the scores.



MODEL COMPARISON

JUDGED ON F1 SCORE

SVM WITH RBF KERNEL BEST OVERALL

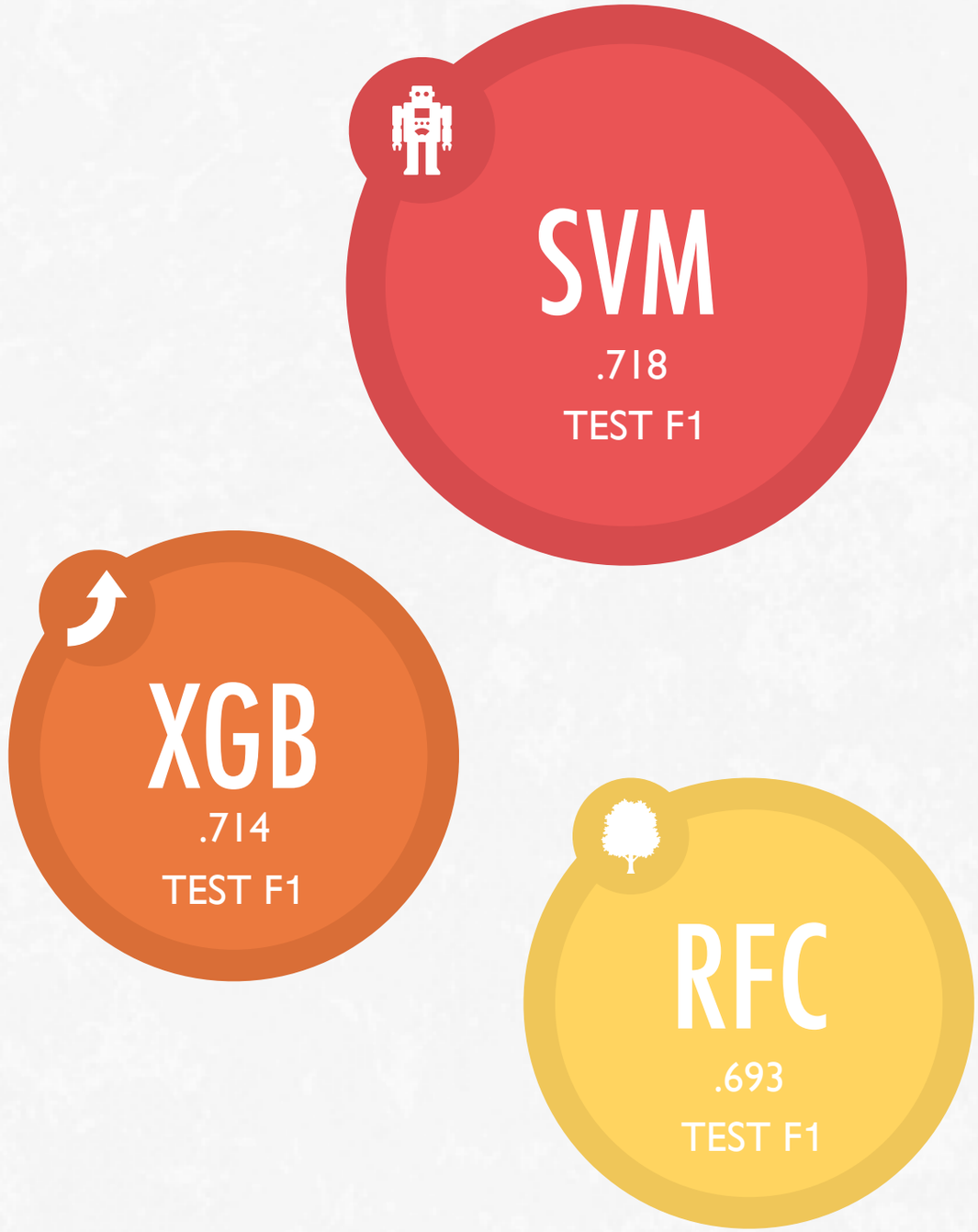
F1 score since care equally about precision + recall.

66-dim features were used: all took advantage of short-time Fourier transform based methods to shrink the original 88,200-dim vectors down.

Essentially a filtering-down effect by:

1. Octave frequency matrix
2. Perceptual pitch matrix
3. Rolloff frequencies

Then the mean AND standard deviation of the above are the features

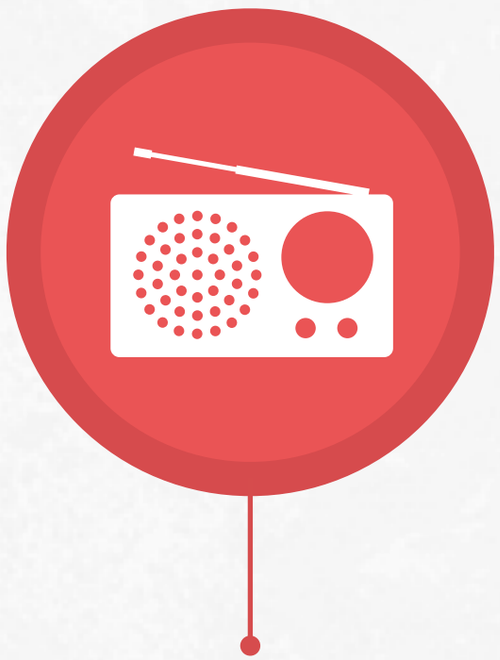


AWS XXL SERVER FOR 10 LOOPS x 9-GROUP CV

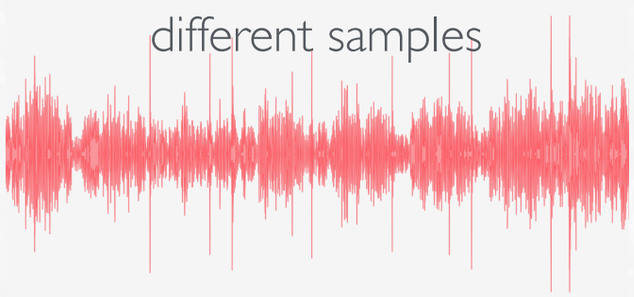
POINTS OF CONFUSION

CLASSES WITH LOW PRECISION AND RECALL

STREET MUSIC

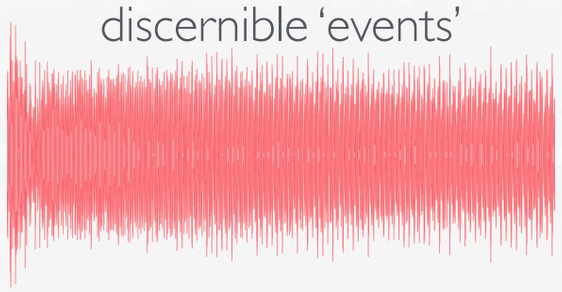


High variation in different samples

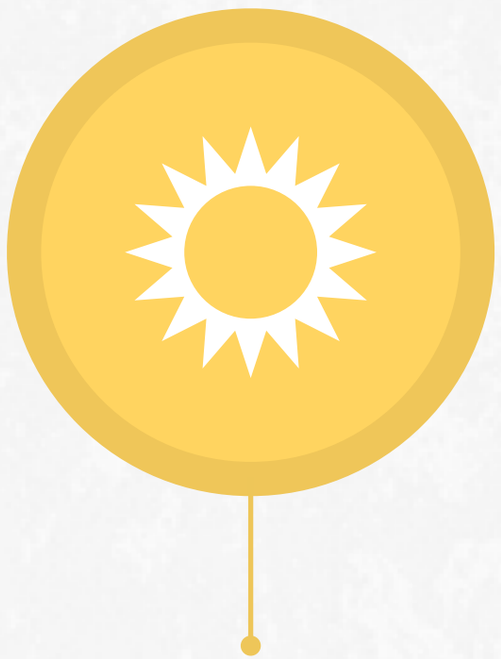


ENGINE IDLING

Very flat rumble, no discernible ‘events’



AIR CONDITIONER



Also a flat rumble, with intermittent spikes

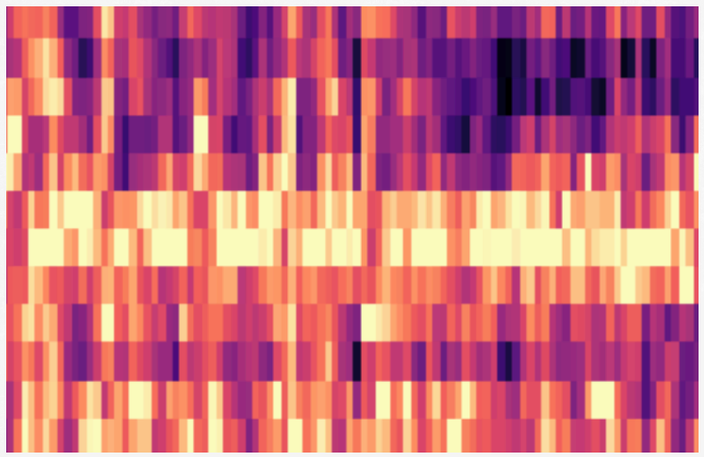
ALL CLASSES: AC, ENGINE IDLING, JACKHAMMER, DOG BARK, CHILDREN PLAYING, STREET MUSIC, SIREN, GUN SHOT, DRILLING, CAR HORN

NEURAL NET - CNN FROM SCRATCH

PYTORCH

CNN ARCHITECTURE

FEEDING IN SPECTRAL PICTURES



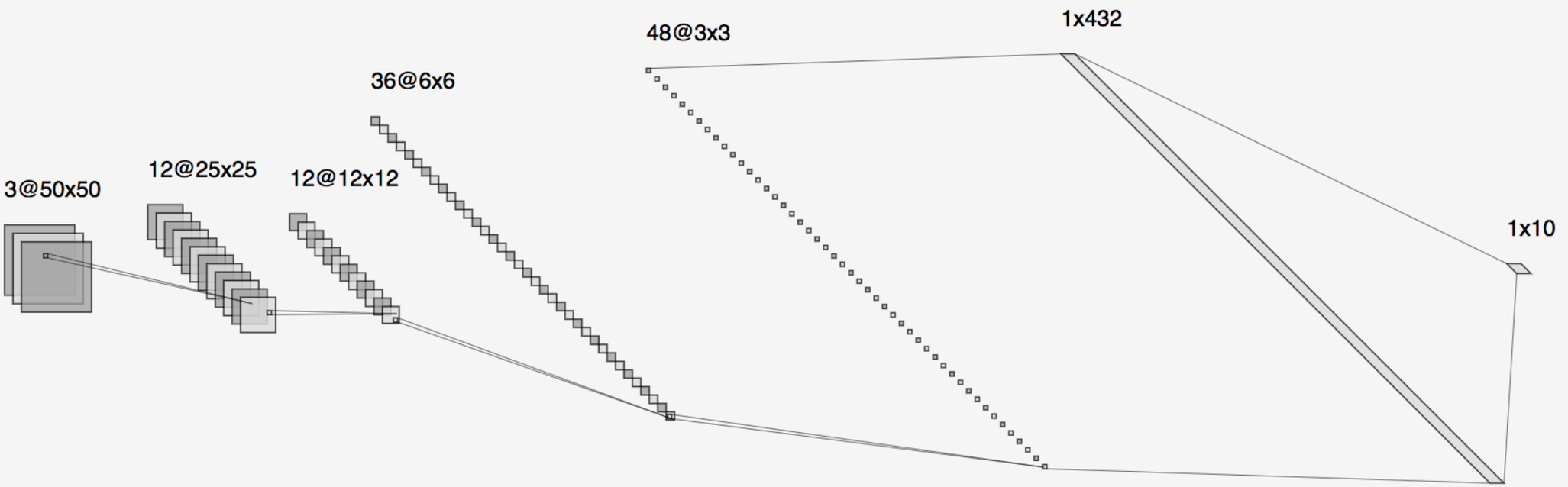
MODEL INPUTS

The images show the frequency intensity of 12 distinct notes across time.

Very quickly overfits the training set despite many layers of regularization.

Future Work: re-create larger and different images to use.

Audio to image is very slow!



THANK YOU

MARC KELECHAVA